引用格式：程学旗, 刘盛华, 张儒清. 大数据分析处理技术新体系的思考. 中国科学院院刊, 2022, 37(1): 60-67.
Cheng X Q, Liu S H, Zhang R Q. Thinking on new system for big data technology. Bulletin of Chinese Academy of Sciences, 2022, 37(1): 60-67. (in Chinese)

大数据分析处理技术新体系的思考

程学旗* 刘盛华 张儒清

1 中国科学院计算技术研究所 北京 100190
2 中国科学院大学 计算机科学与技术学院 北京 100049

摘要 近年来，大数据技术与系统在性能和效率方面已经取得了显著的提升，大数据应用到各个行业，赋能产业智能化发展，成为信息社会进入智能化阶段的关键要素。然而，大数据技术发展也面临着更深层次的挑战，如数据泛滥与高价值数据缺失并存、大数据分析研判复杂不确定、数据流通共享与数据可信安全使用难以兼顾等。这些挑战将推动大数据分析处理技术的创新变革，促进新技术体系的建立与发展。文章面向大数据分析处理面临的新架构、新模式、新范式和安全可信需求，提出构建新一代大数据分析处理系统栈，探索大数据价值利用新范式，并展望新技术体系下的牵引性需求与重大应用。

关键词 大数据，技术架构，系统栈，新模式，新范式，安全可信

DOI 10.16418/j.issn.1000-3045.20211117005

信息社会进入大数据时代后，人们的日常工作和行为、各种在线系统（如信息系统、工业生产线）的工作状态、各类传感器的信号、导航定位系统（全球定位系统 GPS、北斗卫星导航系统等）产生的记录等作为“经验”被常规地记录成为大规模数据。不同于以往为验证科学理论和猜想而记录和收集的科学大数据，记录这些大规模数据起初并没有明确的科学目标。但是，它们却制造了另外的机会。人们可以通过这些数据发现和总结出规律，并依据这些规律提升系统的效率，也可预测、判断未来的趋势，甚至辅助做出更加科学理性的决策[1]。这个过程所依赖的就是大数据分析处理技术。因此，大数据分析处理技术旨在利用数据科学的方法和广泛记录下来的数据，以实现从数据到信息、信息到知识、知识到决策的价值转换[2]。

当前，数字经济成为社会经济的一个重要内涵，数据成为关键生产要素，大数据处理技术越来越深刻地影响着世界的运行状态。随着越来越多的数据被记录、收集和存储，如何深刻洞察数据分布规律、高效挖掘数据价值，成为智能化时代需要解决的关键

*通信作者
资助项目：国家自然科学基金重大研究计划集成项目（91746301），国家自然科学基金面上项目（61772498），国家自然科学基金联合项目（U1911401），国家自然科学基金青年项目（62006218）
修改稿收到日期：2021年12月6日

问题。据美国国际数据公司（IDC）的报告，2020 年全球数据量为 44 ZB 左右，2025 年全球数据量将达到 175 ZB。而这些数据只有 2% 得到了留存，且留存的仅 50% 被使用过①。由此可见，线性提升的数据处理能力并无法匹配指数级增长的数据规模，使得两者之间的“剪刀差”越来越大。与此同时，在庞大的数据空间中，对特定任务真正有价值的核心数据却往往是极度稀疏或不完整的。以上现象即数据泛滥与高价值数据缺失并存的表现。

以互联网平台企业服务为代表的智能化应用大都采用“大数据＋大模型＋大算力”支撑的大数据分析处理技术，主要通过系统的优化来增加数据处理规模并提升计算性能，从而有效解决了一些相对低阶复杂度的预测判定问题，如图像分类、语音识别、结构预测，以及规则明确的人机对弈游戏等。而在开放复杂的系统环境下，数据动态生成演化，影响系统运行状态的不确定因素和变量巨大，从而导致一些高阶复杂的问题难以直接模型化，或近似求解的结果不可信，如金融风险预测、个性智能诊疗、开放环境下的自动驾驶等。在这些高阶复杂的真实系统中，数据采集分布往往是不稳定和不完备的，这对要求精准判别的大数据分析处理模式提出了新的挑战。

同时，解决数据与算法的安全可信问题也已迫在眉睫。数据流通共享的过程中面临着数据滥用、隐私泄露的情况。数据本身可能也会引入真实世界存在的偏差，或者在对抗攻击下数据被污染，使得大数据分析模型做出有偏的、错误的决策[3]。在大数据分析处理技术逐渐应用于关键领域的当下，如何让大数据技术以一种安全可信的方式服务于各个领域，是未来大数据发展必须面对的又一个难题。

本文首先回顾了近 10 年来大数据技术的发展现状，并针对数据泛滥与数据缺失并存、大数据分析研判的复杂不确定性和数据安全缺失等挑战，提出大数据分析的新范式和安全可信的大数据处理新架构，探索大数据支撑智能应用的新模式。在此基础上，提出构建新一代大数据分析处理软件栈，并展望新技术体系下的牵引性需求与重大应用。

1 大数据分析处理现状

近 10 年来，在产学研各界及政府主导的大力推动下，大数据技术架构、生态环境及各行各业的大数据应用发展迅速。

1.1 大数据技术架构

*海量数据促进了大数据技术架构的发展。① 大数据管理技术方面。*传统关系数据库（SQL）主要处理较少数据和较小并发访问规模，而且存在大量读写硬盘和日志记录操作，难以横向扩展，无法满足互联网应用的数据管理需求。为了实现更多的数据管理、更大规模的并发访问及更多样的数据模式，面向特定需求的各类非关系型数据库（NoSQL）和从底层重构的分布式关系数据库（NewSQL）正在快速发展[4]。其中，NewSQL 保持了传统数据库支持事务处理正确执行四要素（ACID）②和 SQL 标准查询等特性，并具备与 NoSQL 同样优秀的可扩展性。*② 大数据处理技术方面。*根据处理需求的不同，存在多种不同的并行计算模型，包括以 Hadoop、Spark 为代表的批处理，以 Spark Streaming、Flink、STORM 为代表的高实时性的流处理，以 Apache Beam、Lambda 为代表的流批一体混合处理[5]，以及以 GraphX、Apache Giraph 为代表的图处理③。同时，图数据和实时数据处理的爆发

① Reinsel D, Gantz J, Rydning J. Data Age 2025: The Evolution of Data to Life-Critical, IDC White Paper. USA: IDC, 2017.

② 指数据库管理系统（DBMS）在写入或更新资料的过程中，为保证事务（transaction）是正确可靠的，所必须具备的 4 个特性：原子性（atomicity，或称不可分割性）、一致性（consistency）、隔离性（isolation，又称独立性）、持久性（durability）。

③ Gonzalez J E, Xin R S, Dave A, et al. Graphx: Graph processing in a distributed dataflow framework. (2014-10-06)[2021-12-31]. https://dl.acm.org/doi/abs/10.5555/2685048.2685096.

性需求也推动了图流处理模式的融合[6]。除此之外，计算硬件逐渐发展为多种计算单元（如 CPU、GPU、NPU 等）组成的异构计算系统，新型硬件和软件的多层次融合进一步提升了大数据处理效率。③ 大数据分析技术方面。分析需求逐渐从小规模、单源、单一模态数据的统计挖掘分析转变为海量、多源、多模态数据的复杂异质关联。深度学习技术的快速发展，推动了大数据分析模型能力的提升。神经网络模型在 2012 年的计算机视觉的目标识别项目 ImageNet 比赛夺冠后重回人们的视野，随后诞生了一系列突破性的工作，包括知识图谱提供知识服务、生成对抗网络合成真实数据、AlphaGo 围棋战胜人类、GPT-3 预训练语言模型等。此外，日益成熟的深度学习框架（如 TensorFlow、PyTorch、飞桨等）也降低了使用深度学习分析大数据的门槛。

1.2 大数据应用

近年来大数据分析处理技术飞速发展，催生了众多大数据应用，赋能了大量行业的智能化发展，一些标志性的应用从模式和能力上颠覆了传统的信息技术能力。① 科学发现方面。DeepMind 公司的 Alphafold 可基于蛋白质的基因序列数据预测蛋白质的三维结构，进而分析蛋白质的属性，帮助生物学取得了重大进展[7]。② 数字经济方面。电商平台的兴起，连接遍布全球各个角落的消费者和供货方，通过交易大数据的精准分析，提高了交易效率，推动了在线支付与数字货币的使用，颠覆了社会征信的模式；基于大数据进行的金融风险研判、小微金融和普惠式金融等也促进了数字经济的繁荣。③ 社会安全方面。我国使用大数据方法辅助公共卫生、金融等领域的社会治理与决策；美国尝试研究大数据技术在解决社会不平等、城市政策制定方面的作用。④ 生命健康方面。英国基于海量学术论文和临床试验结果研发了治愈运动神经衰退等多种药物④，以及近两年各国大量使用的数字接触追踪技术，辅助预测了疫情传播速度和趋势⑤，分别被列入《麻省理工科技评论》2020 年和 2021 年的“全球十大突破性技术”。国内外大数据技术的应用改变了诸多传统行业中耗时耗力的工作方式，取得了智能高效的丰硕成果。

1.3 大数据生态建设

大数据分析处理的繁荣离不开大规模数据资源共享、技术架构开放和算法模型开源所形成的技术生态发展。① 开源数据方面。开源数据支撑各类大数据技术的构建。例如，2009 年美国斯坦福大学发布的视觉数据集 ImageNet[8]、2015 年美国麻省理工学院发布的大规模医疗信息数据库 MIMIC-Ⅲ[9]、2020 年斯坦福大学发布的图数据集 Open Graph Benchmark[10]，都极大地影响了大数据技术的发展。② 开源软件方面。Apache 软件基金会基于 Hadoop 生态先后发布了一整套完善的分布式存储与处理框架 Map-Reduce、线性代数计算框架 Mahout、机器学习库 MLlib 等，旨在让开发者快速实现和应用大数据分析处理算法。2014 年以来，深度神经网络的开源框架，如 Caffe、Tensorflow、PyTorch 等，更是为从大数据中学习面向不同任务的智能模型提供了重要支持。③ 开源模型方面。基于大规模数据学习的 BERT、GPT3 等预训练语言模型[11]，大幅降低了相关技术的应用成本，拓宽了下游应用场景。此外，如何保障数据安全和个人隐私，最近也得到了各国政府和组织的高度重视。因此，兼顾技术发展和数据安全，平衡效率和风险，建立良好的大数据生态环境，仍需要进一步探索。

2 新一代大数据分析处理需求

当前针对大规模异质化数据集合，主流的大数据分析处理方法是在通用模型框架下不断尝试超大规模

④ https://www.technologyreview.com/10-breakthrough-technologies/2020.
⑤ https://www.technologyreview.com/2021/02/24/1014369/10-breakthrough-technologies-2021.

的模型参数，实现“端到端”的分析推断。在这种模式下，大数据分析处理能力很大程度依赖于算力平台和数据资源的支持。在实际应用中，这些大数据分析处理技术面临着真实场景和关键领域中数据泛滥与缺失并存、大数据分析研判的复杂不确定性、数据安全监管缺失等挑战，最终使得分析处理存在过程可解释性差、模型泛化能力弱、因果规律不清晰、研判结果不可信、数据价值利用率低等问题。为解决这些挑战性问题，我们需要重新思考大数据处理架构与分析模式，新一代的大数据分析处理技术体系应该在各种实时场景下实现高价值知识生成、持续在线的瞬时决策、安全可信的推理研判，以及适用于未来各种有人-无人结合的在线系统行动优化。本文认为，新一代大数据分析处理至少需要满足如下4个方面的需求。

（1）*人在回路的计算范式*。为解决现有大数据分析处理方法难以攻克的高阶复杂问题，需要在其中引入人的智能与决策，强调人、机器及数据之间的有机交互。不同于原来的人机交互，即机器按照人的指令，或人听机器的输出结果，而是更关注人脑和机器思维的深度融合计算[12]。

（2）*广谱关联的分析模式*。为解决大数据价值密度低、极稀疏、不均匀、关键信息缺失的问题，一方面，融合各个对象在“人机物”融合的多域多维数据空间中留下的多元异构信号，利用关联增强信号；另一方面，融合数据与知识，构建终生学习、可迁移扩展的知识体系，形成数据驱动与知识制导深度融合的新分析模式。

（3）*在线增强的处理架构*。随着万物互联和智能泛在发展，大数据云边端协同计算技术和解耦化的云边端处理框架成为热点。基于云计算环境下的流批混合处理将进一步向边缘端发展，训练学习与推理预测将在前端设备上融合一体。利用云边端资源弹性调度能力，实现感知与认知能力前置，支持在线环境下基于动态活性数据的瞬时决策，从而形成去中心化、异构分布、持续在线的新型计算框架。

（4）*安全可信的大数据分析*。安全可信是满足关键领域和场景下认知和决策安全的基本需求。一方面，着重关注大数据分析处理结果的可解释、可信和公平性[3,13]；另一方面，实现数据在收集、存储、使用、流通中的安全保护和异常检测，保证在强对抗攻击下分析处理模型与方法的鲁棒性和免疫性。

3 新一代大数据分析处理软件栈

在高效的大数据价值提取、安全可信的分析处理目标下，针对以上4个大数据分析处理的重要需求，未来急需建立自立自强的大数据分析处理技术新体系，发展新一代大数据分析处理软件栈（图1），从底层数据操作系统、通用分析处理中间件、业务驱动的计算环境及框架3个方面进行研究。

3.1 全栈式的大数据系统软件

发展并涵盖数据接入、流式处理、图计算、训推一体⑥等多个方面的大数据系统软件。

（1）*数据接入方面*。针对当前数据采集流程中数据来源繁多、数据类型混合及异质数据存储效率低下的难题，研究“人机物”融合的数据汇聚与融合方法，支持对多种数据源的结构化、半结构化数据的采集与融合，探索高效的存储算法，提高底层存储空间利用效率，支持对数据的高效压缩与还原，实现对“人机物”三元数据空间中的多源异构数据进行高效感知、采集、融合与存储，为系统提供高质量的数据流接入。

（2）*流式处理方面*。现有大数据处理框架中存在计算模式单一的问题，即单独追求大批量或强时效。针对这一问题，将研究多计算模式融合的流式处理框

⑥ 在人工智能深度学习中指训练和推理一体化。

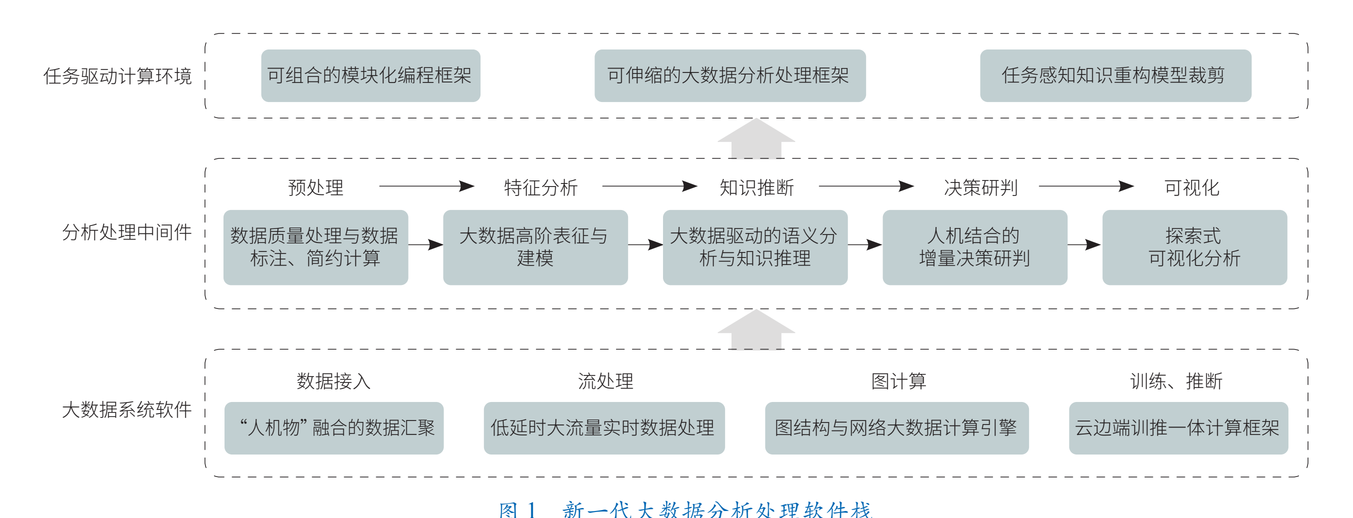


图 1　新一代大数据分析处理软件栈

Figure 1　New big data analyzing and processing system stack

架，支持批处理、流处理、图处理等多种计算模式，实现低延时、大流量、强时效的数据处理，以应对不断接入的高速数据流。

（3）关联数据的计算方面。现有计算框架难以适应图结构数据的强数据依赖性、高随机访存与非均匀幂律分布特性。针对这一问题，研究针对图结构和网络大数据的计算引擎，提出大规模图数据的新型分布式计算框架和并行计算机制，定制大规模图数据的查询语言标准与规范，实现图查询与图分析语言的标准化。

（4）训练推理方面。现有云端大数据处理架构难以满足大规模服务的实时性与计算资源需求。针对这一问题，研究云边端协同的训推一体框架，将大数据分析处理中的训练与推断流程从云端推向边缘，支持训推一体[14]，在数据生成的边端提供服务和执行计算，实现“认知前置”和终生学习，以提供分布式、低延迟、持续在线的智能服务和瞬时决策。

3.2 重构大数据分析处理流程

从预处理、数据表征、语义分析与知识推理、决策研判到可视化的全技术链上升级创新。

（1）数据质量处理与简约计算方面。针对数据质量处理，可发展利用群智技术挖掘高质量数据，以低成本、高效率的方式实现大规模数据的采集处理；针对简约计算方面，可研究基于数据复杂度的近似计算理论和优化算法框架，以此指导人们寻找面向计算的数据内核或者数据边界的基本方法，构建具有高效计算能力的模型。

（2）大数据高阶表征与建模方面。探索基于无监督预训练的数据表征学习的理论与方法，从大规模未标注的语料数据中抽取高层次语义抽象的数据表征，提高语义表征的泛化能力；研究基于小样本数据的预训练—微调模型，在大规模无监督语料训练得到的数据表征基础上，构建辅助上层任务的通用高质量数据表征；探索基于领域知识的预训练数据建模理论与方法，将人类知识融合到预训练模型中，提升预训练模型的学习效率等。同时，为应对数据多源异构造成的知识隔阂，有必要进一步发展跨模态数据表征和建模、多源知识融合技术，以实现全域知识联合和利用。

（3）大数据驱动的语义分析与知识推理方面。研究面向细粒度语义单元的大数据语义融合方法，显著提高多源异构数据关联融合的效果；研究样本稀疏环境下的领域知识获取、大规模常识获取与理解、知识获取中的人机协作机制与方法，提升知识获取的能

力，大幅提高知识库的规模；研究基于知识图谱的可解释分析方法、数据驱动与知识引导深度融合的新型语义分析方法，显著提升知识驱动下各类模型的效果和可解释性。

（4）人机结合的增量决策研判方面。未来大量物理设备、无人设备、人脑，通过泛在网络实现“上线”和“互联”，为人的参与提供了基本的物质条件。人作为具备智能的自然系统，如何参与到机器智能的系统回路中是一个关键问题。未来应重点解决思维融合或决策融合的问题，探索人脑数据及机器智能系统信息可相互转换的新型数据科学理论，并设计高效能的计算方法。当下的算法模型不会随着数据的生成而持续学习，即无法应对连续和意外变化的环境，特别是在任务关键型应用程序中更需谨慎。因此，研究持续学习、在线学习等技术，实现算法模型持续在线瞬时决策十分必要。

（5）探索式可视化分析方面。研究新型的跨主体（人、机、物）可视交互理论，构建多人协同的混合主动式可视分析范式，支持多人同时对相同或不同的可视化视图进行多角度的探索，设计相应的可视表达与交互形式；研究围绕大数据可视化的认知计算与聚合理解模型、方法与核心技术，构建人机协同智能及其驱动的大数据可视内容与属性的自动理解关键技术；提升围绕大数据可视化的计算机自动理解、表示与生成能力等，构建大数据可视计算与交互技术体系。

3.3 建立任务驱动的大数据计算环境

从可组合的模块化编程框架、可伸缩的大数据分析处理框架、任务感知的知识重构模型裁剪这 3 个方面发力，为各行各业提供场景感知、共识感知的更优质和更灵活的分析处理环境。

（1）可组合的模块化编程框架方面。未来可发展面向多业务可扩展、可重构的敏捷开发框架，构建多形态分析模式库和智能业务编程框架，突破多源异构数据的关联分析和全息展示，实现对数据、算法、模型的高层次抽象，形成支撑面向任务场景的智能组合分析算子库，实现智能算法的内生性支持，赋能人机混合的交互式协同分析。

（2）可伸缩的大数据分析处理框架方面。未来可发展支持弹性计算、可伸缩模型、可弹性配置的处理框架，即根据实际应用的任务场景与计算资源的需求等方面划分各种任务，满足特定需求、精度需求、延时需求、实时性需求等；同时，构建可伸缩的大数据分析处理框架，能够灵活配置计算资源和数据规模，以实现弹性适配。

（3）任务感知的知识重构和模型裁剪方面。未来可发展面向任务的高级知识计算语言和模型裁剪技术，基于通用知识图谱实现面向特定领域任务的知识重构，建立起常识与领域知识融合的知识计算引擎，显著提升知识管理和利用的能力与效率。

4 推动新一代大数据分析处理技术发展建议

（1）建立理论基础。大数据分析处理技术新体系的建立，离不开基础理论的突破。① 建立数据复杂性和大数据可计算性理论。回归数据本原，探索数据在分布规律、结构规则和时空尺度方面的规律性，以此设计高效能的计算方法。② 探索异质广谱关联的大数据分析理论。将各类目标在“人机物”融合的多维数据空间留下的微弱信号进行关联放大，研究广域开环、非统一量纲环境下瞬时决策推断方法的收敛性理论。③ 研究大数据分析处理的安全可信理论。一方面，研究数据的安全共享和隐私计算理论，保障数据流通共享过程中的安全性；另一方面，研究数据的固有偏差性和数据遭受攻击时的分析处理的鲁棒性极限和可验证理论，建立可防范、可审计、可追责的机制，保证强对抗环境下分析处理结果的可信。

（2）加大应用牵引。新大数据分析处理技术体系应能全面高效赋能行业、产业、安全领域。同时，

还需要利用科学发现、生命健康、社会治理等牵引性应用场景来推动大数据分析处理新体系的健康、良性发展。① 科学发现方面。研究借助大数据分析技术从大量实验数据中发现科学规律，形成基于大数据分析的新型科学研究方法论。② 生命健康方面。研究大数据方法用于辅助复杂化合物分子的发现，降低新型药物的研发成本，加快提升综合医疗水平，利用大数据手段应对重大疫情和事件的高效用、持续在线决策。③ 社会治理方面。充分发挥大数据技术在多方复杂关联问题、社会群体认知建模分析中的优势，构建人工辅助智能决策系统，实现政府决策科学化、社会治理精准化、公共服务高效化。

（3）数据治理生态环境。大数据技术的应用与发展离不开良性的数据治理和技术生态建设。① 个人隐私保护。需要相应的法律法规加以规范。例如，欧盟2016年出台了《通用数据保护条例》，帮助公民控制个人隐私数据；我国于 2021 年发布了《中华人民共和国数据安全法》和《中华人民共和国个人信息保护法》，对大数据的采集与使用给予合理的管控和监督。② 保证数据的安全流通共享。需要建立数据流通交易规则规范，优化数据共享、交易、流通相关的制度，明确数据权属分配，探索数据交易市场，构建有序的数据流通环境。

5 结语

综上所述，未来应发展突破通用模型架构、分析模式和计算范式，建立新架构、新模式、新范式，以及安全可信的大数据分析处理技术新体系；构建新一代大数据分析处理软件栈；研究和发展相应的理论，践行牵引性应用；建立良性的数据治理生态，推动大数据分析处理技术的持续进步和跨越式发展。

参考文献

1 李国杰, 程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考. 中国科学院院刊, 2012, (6): 647-657.
Li G J, Cheng X Q. Research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences, 2012, (6): 647-657. (in Chinese)

2 徐宗本, 唐年胜, 程学旗. 数据科学：它的内涵、方法、意义与发展. 北京: 科学出版社, 2021.
Xu Z B, Tang N S, Cheng X Q. Data Science: Its concept, method, meaning and development. Beijing: Science Press, 2021. (in Chinese)

3 Wing J M. Trustworthy AI. Communications of the ACM, 2020, 64(10): 64-71.

4 梅宏. 大数据导论. 北京: 高等教育出版社, 2018.
Mei H. Introduction to Big Data. Beijing: Higher Education Press, 2018. (in Chinese)

5 de Assuncao M D, Da Silva Veith A, Buyya R. Distributed data stream processing and edge computing: A survey on resource elasticity and future directions. Journal of Network and Computer Applications, 2018, 103: 1-17.

6 Mcgregor A. Graph stream algorithms: A survey. ACM SIGMOD Record, 2014, 43(1): 9-20.

7 Senior A W, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature, 2020, 577: 706-710.

8 Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. IEEE Computer Vision and Pattern Recognition, 2009, (1): 248-255.

9 Johnson A E W, Pollard T J, Shen L, et al. MIMIC-III, a freely accessible critical care database. Scientific data, 2016, 3(1): 1-9.

10 Hu W, Fey M, Zitnik M, et al. Open Graph Benchmark: Datasets for Machine Learning on Graphs. Neural Information Processing Systems (NeurIPS), 2020.

11 Qiu X, Sun T, Xu Y, et al. Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 2020, 63(10): 1-26.

12 程学旗, 梅宏, 赵伟, 等. 数据科学与计算智能：内涵、范式与机遇. 中国科学院院刊, 2020, 35(12): 1470-1481.
Cheng X Q, Mei H, Zhao W, et al. Data science and

computing intelligence: Concept, paradigm, and opportunities. Bulletin of Chinese Academy of Sciences, 2020, 35(12): 1470-1481. (in Chinese)

13 Kleinberg J, Ludwig J, Mullainathan S, et al. Discrimination in the Age of Algorithms. Journal of Legal Analysis, 2018, 10: 113-174.

14 Stoica I, Song D, Popa R A, et al. A Berkeley View of Systems Challenges for AI. 2017. (2017-10-16)[2021-11-10]. http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.pdf.

Thinking on New System for Big Data Technology

CHENG Xueqi* LIU Shenghua ZHANG Ruqing

(1 Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;

2 School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract In recent years, there are such significant improvements on the performance and efficiency of big data technology and system. As it is widely applied in various fields, big data has empowered industrial intelligence, and is the key step into the intelligent stage of information society. Therefore, we are facing greater challenges nowadays, such as the paradox of data flooding and high-value data lacking, the complexity and uncertainty of big data analysis, and the difficulty to balance the data on sharing and circulation, and trustworthiness and security. Moreover, these challenges will not only promote the innovation and change of big data technology, but also develop and establish a new technology system. With respect to the requirements of new architecture, new paradigm, new model and security and trustworthiness, this study proposes to build a new big data analyzing and processing system stack, explore the new paradigm of extracting big data value, and outlook on the pioneer applications as the traction to a broad range of fields.

Keywords big data, technology architecture, software system stack, new model, new paradigm, security and trustworthy



程学旗 中国科学院计算技术研究所副所长、研究员，国家杰出青年科学基金、国务院特殊津贴获得者。担任中国计算机学会（CCF）大数据专家委员会秘书长、中国工业与应用数学学会（CSIAM）大数据与人工智能专委会副主任、中国中文信息学会信息检索专委会主任。在大数据分析系统、 Web信息检索与数据挖掘等领域发表学术论文200余篇，获授权发明专利60余项。获国家科技进步奖二等奖3项、国家技术发明奖二等奖1项。E-mail: cxq@ict.ac.cn

CHENG Xueqi Professor and Deputy Director of the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). He was funded by the National Science Fund for Distinguished Young Scholars of National Natural Science Foundation of China and enjoys the State Council Special Allowance. He is the Secretary General of the CCF Task Force on Big Data, the Deputy Chair of the CSIAM Activity Group on Big Data and Artificial Intelligence, and the Chair of the Technical Committee on Information Retrieval, Chinese Information Processing Society of China. He has published more than 200 papers, and has more than 60 authorized patents on big data analysis system, web information retrieval and data mining. He has awarded three times of second prize of National Science and Technology Progress Award and one second prize of National Technology Invention Award. E-mail: cxq@ict.ac.cn

■责任编辑：文彦杰

*Corresponding author